Menemui Matematik
(Discovering Mathematics)

journal homepage: https://persama.org.my/dismath/home

Statistical Power of Model Selection Methods in Extreme Value Modelling

Sunday Samuel Bako[1,2], Norhaslinda Ali[1,3*] and Jayanthi Arasan[1,3]

[1]*Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, 43400 UPM Serdang, Selangor*
[2]*Department of Mathematical Sciences, Kaduna State University, 800241 Kaduna, Nigeria*
[3]*Department of Mathematics and Statistics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor*

[*]norhaslinda@upm.edu.my
*Corresponding author

Received:13 Jun 2025
Accepted:13 August 2025

ABSTRACT

The objective of model selection in extreme value modelling is to identifying the probability distribution that best fits an observed sample. Statistical criteria are commonly used for this purpose; however, they often have limitations, particularly in distinguishing between distributions with similar tail behaviour, especially when the sample size is small and the distributions are asymmetric. In this study, we demonstrate, using trimming and subsampling techniques, the ability of model selection methods to reject a candidate distribution when it differs from the true underlying distribution. Four model selection methods are examined: Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC), and the Anderson Darling test. Results from the power analysis show that BIC is the most effective method for identifying the Lognormal and Gumbel distributions, while AIC performs best for the Pearson Type III (P3) distribution. In terms of the comprehensive power, AIC demonstrates the highest power, followed by AICc, BIC, and the Anderson Darling test. These findings demonstrate that the use of trimming, subsampling, and appropriate model selection methods is a viable technique for distinguishing among candidate distributions and evaluating the power of model selection criteria. This approach provides a practical framework for more reliable model selection in extreme value analysis.

Keywords: Extreme value, Model selection, Trimming, Subsampling

INTRODUCTION

Extreme value modeling is an important tool for understanding and predicting rare events across various fields, including environmental sciences and finance (see, for example, Kousar et al., 2020; Chan et al., 2022). In recent times, the increasing rate of extreme events such as rainfall has posed a significant threat due to global warming scenarios, leading to constant flooding caused by heavy rainfall, which results in huge losses of human life and property globally. This has led to debates in the public domain about the apparent causes of increasing extreme events. Researchers have queried the varying rainfall intensity statistics under greenhouse conditions. The theory of extreme values and their distributional models are often employed to accurately represent the rate of these extreme events.

In extreme value analysis, selecting an appropriate probability distribution is fundamental to accurately represent the behavior of extreme events. Commonly used distributions include the Generalized Extreme Value (GEV), Lognormal, Gumbel, and Pearson Type III (P3) distributions (Cunanne 1989). Model se- lection criteria such as the Akaike Information Criterion (AIC), corrected AIC (AICc), Bayesian Information Criterion (BIC), and the Anderson-Darling (AD) test are frequently employed to identify the best-fitting distribution for a given dataset. Laio et al. (2009) found that AIC and BIC are similarly effective for selecting extreme value distributions and introduced an Anderson-Darling based criterion for model selection. They recommend using either AIC or BIC alongside the AD test; if both agree, the selected model is reliable. If not, the discrepancy reflects equifinality. Building on this, Di Baldassarre et al. (2009) evaluated AIC, BIC, and the AD test for quantile estimation, noting similar overall performance, with the AD test performing better as L-skewness increases. These studies highlight the potential of these criteria to reduce un- certainty in extreme value modeling.

Despite the widespread use of these model selection methods, they often face challenges in distinguishing between distributions with similar tail behaviors, especially when dealing with small sample sizes or asymmetric distributions. This limitation can lead to the selection of suboptimal models, thereby affecting the reliability of predictions and subsequent decisions based on the analysis (Beirlant & Bladt, 2025; Brewer et al., 2016).While some studies have examined the statistical power of model selection methods, particularly their ability to correctly reject incorrect models in favor of the true underlying distribution (Zeng et al., 2015; Reghenzani et al., 2019), the potential of integrating data preprocessing techniques like trimming and subsampling to improve their discriminatory power remain largely unexplored.

Resampling methods like the bootstrap (Efron, 1979) and subsampling (Politis et al., 1999) are widely recognized, data-driven simulation techniques employed for statistical inference. Subsampling generate multiple datasets from the original sample, typically of fixed but smaller size. Unlike the bootstrap, subsampling selects observations without replacement. A study by Politis (1994) and Chernick (2011) has demonstrated that subsampling can produce consistent estimators. According to Politis (1994), its key advantage lies in generating resamples that closely reflect the underlying distribution, making it a valuable tool for distribution selection. Trimming and censoring are used in extreme value analysis to reduce the influence of smaller observations on the upper tail without relying on unrealistic assumptions about the data generating process, which is critical in modeling extreme events (Bhattarai, 2004).

This study aims to evaluate the statistical power of four widely used model selection methods: AIC, AICc, BIC, and the Anderson-Darling test, in the context of extreme value modeling. By employing trimming and subsampling techniques, we assess the ability of these criteria to correctly identify the true underlying distribution and their ability to correctly reject incorrect models in favor of the true underlying distribution across different scenarios and sample sizes. The primary objective of this paper is to provide a comprehensive analysis of the effectiveness of model selection methods in extreme value modeling, with a focus on their statistical power. By exploring the impact of trimming and subsampling, we aim to offer practical insights and recommendations for researchers and practitioners seeking to enhance model selection accuracy in the analysis of extreme events.

METHODOLOGY

Selection of Candidate Probability Distributions

The reports by Cunnane (1989) summarize probability distributions commonly used in extreme value analysis worldwide. Both peak over threshold and annual maximum data methods are applicable, but the former is less common due to challenges in threshold selection and method complexity (Pan and Rahman, 2021). While annual maximum stream flow data can provide reliable estimates, annual maximum rainfall observations are often preferred for their spatial and temporal availability in modeling extreme events (Flammini et al., 2022). This study selects seven widely recommended probability models for analysis of extreme values: Gumbel, lognormal, Frechet, Pearson type III (P3), generalized extreme value, normal, and log-Pearson type III (Cunnane, 1989).This study employs the maximum likelihood estimation (MLE) method to estimate the parameters of the distribution. A key advantage of adopting the MLE approach is its asymptotic efficiency. Moreover, standard and widely applicable approximations are available for a variety of important sampling distributions (Coles, 2001). The following section provides a brief overview of the maximum likelihood method.

Selection of Candidate Parameter Estimation Methods

The maximum likelihood estimation (MLE) method estimates distribution parameters by maximizing the log-likelihood function based on a set of independent and identically distributed observations. Given a probability density functionf(x_i; θ), where θ denotes the vector of unknown parameters, the log-likelihood is defined as $l(\theta) = \sum_{i=1}^{n} log\, f(x_i; \theta)$. The parameter estimates are those that maximize this function, typically using numerical optimization techniques. Although computationally intensive, particularly for models with multiple parameters, MLE generally provides estimators with desirable statistical properties (Coles, 2001; Gado, 2016; Kobierska et al., 2018; Haddad and Rahman, 2011).

Model Selection Methods

To evaluate the appropriateness of different probability distributions in the context of subsampling and trimming, this study considers the Akaike information criterion (AIC), the corrected Akaike information criterion (AICc), the Bayesian information criterion (BIC), and the Anderson-Darling (AD) goodness-of-fit test. Additionally, the Kullback-Leibler information measures the difference between the true model $f(y)$ and a model that approximates it more closely,$K_j = g_j(y, \hat{\theta})$, is utilized by the AIC (Akaike, 1998). The AIC is expressed as follows:

$$-2\, ln[L(D \mid \hat{\theta})] + 2m \quad (1)$$

where m represents the total count of parameters derived for estimation for the j^{th} underlying model, and $L(D \mid \hat{\theta}) = \prod_{i=1}^{n} g_j(y_i, \hat{\theta})$ is the likelihood function (Linhart and Zucchini, 1986). For model selection, the maxima of the log-likelihood function are used, with a larger penalty applied on models that have more estimated parameters m. In practice, after calculating the AIC_j, the best-fitting model is identified as the one with the least AIC. A second-order version of AIC, known as AICc, is defined as

$$-2\, ln[L(D|\hat{\theta})] + 2m\,(\frac{n}{n} - m - 1) \quad (2)$$

with n representing the sample size. The distinction between AIC and AICc lies in the fact that AICc applies a more substantial penalty for the number of parameters estimated m compared to AIC (Calenda et al., 2009; Burnham and Anderson, 2004). The BIC is like the AIC but is based on a Bayesian approach. The concept of the BIC was introduced by Schwarz (1978) and its formula is expressed as

$$-2\ ln[L(D \mid \widehat{\theta})]\ +\ m\ ln(n). \tag{3}$$

The BIC applies a heavier penalty for the number of estimated parameters m that is greater than the AIC. The Anderson-Darling test evaluates an observed cumulative distribution function (CDF) against a theoretical CDF, placing more emphasis on the tails of the distribution. The test statistic for the AD test can be represented as follows:

$$-n\ -\ (1/n)\ \sum_{i=1}^{n}\ (2i\ -\ 1)\ [ln\ F(y_i)\ +\ ln(1\ -\ F(y\_(n-i+1)))]. \tag{4}$$

By emphasizing the tails, the Anderson-Darling test measures model fit through a weighted sum of squared differences between sample and theoretical distributions.

TEST PROCEDURE

According to Cunnane (1989), the Gumbel, lognormal and Pearson Type III (P3) distributions, which are widely used two-parameter and three-parameter models in extreme value analysis, will serve as the parent distributions for the simulation experiment. Multiple subsamples will be drawn from these distributions without replacement, and each subsample's length is based on the size of the initial sample. Trimming improves the representation of maxima in the data (Wang, 1996), and the trimming proportions are carefully selected to retain a sufficient number of observations for analysis. As noted by Politis et al. (1999), subsampling requires that each subsample be smaller than the original dataset. The AIC, AICc, and BIC model selection techniques are applied to both trimmed and untrimmed subsamples, selecting the distribution that yields the smallest value. The Anderson-Darling (AD) test is also used to evaluate the goodness of fit. This procedure is repeated for various subsample sizes, and the selection frequency of each candidate distribution is recorded. The procedure for the power test is structured as follows:

1. Let $f(y)\ =\ g_j\ *\ (y, \theta\ *)$ represent the parent distribution with specified parameters. Generate a substantial number of subsamples, each of size$n$, from the parent distribution.

2. A subsamples of size b, where $b\ <\ n$, is chosen without replacement from the subsamples generated in step 1. The subsample size b is determined by the size n of the original sample.

3. Trimming is applied to each subsamples of size b that is chosen. The trimming proportions will be systematically selected so as not to take away too many observations, thereby reducing the number of samples available for analysis.

4. Fit the candidate distributions, $j\ =\ 1, 2, 3, .., 7$, of the trimmed and untrimmed (0%) samples in step 3 to the model selection methods, and compute their values, AIC_j, $AICc_j$, BIC_j and AD_j, for each of the candidate distributions.

5. The preferred model K_i* is the one with the smallest AIC score, denoted as $AIC_i* = AIC\_min$. If $i \not\equiv$ equals $j*$, then the AIC is chosen because it accurately identifies the actual underlying distribution. Similar procedure is carried out for AICc and BIC. In the AD test, if the non-exceedance probability $P(A_2)$ of the test statistic A^2 exceeds$k$, where$k = 1 - \alpha$, the candidate distribution is rejected.

6. The testing procedure in steps 1−5 is repeated for different subsample sizes b, and the number of times each selection method successfully identifies the optimal model is recorded.

The acceptance proportion for each parent distribution is defined as follows:

$$PMSH_{bi} = MSH_{(bi,j)} / b \tag{5}$$

where M is the method used for model selection, S is the function used as parent distribution, H is the candidate distribution, b is the total subsample size used, b_i is the ith sample of the subsample b, j is the trimming proportion, and $MSH_(b_{i,j})$ is the accepting times for a specific candidate distribution.

The samples with length n with subsample size b were generated from the Lognormal, Gumbel, and P3 distributions, and the Normal, Gumbel, EV2, GEV, P3, LP3, and Lognormal distribution are used as candidate distributions. After that, we estimate the parameters of the samples for all the candidate distribution using the maximum likelihood estimation method for both the trimmed and untrimmed samples, and test for the most suitable models amongst the candidate distributions using the model selection methods; AIC, AICc, BIC, and the AD tests. The length of the data and subsample size was varied. The subsample size depends on the length of the data. The significance level of the AIC, AICc, BIC, and AD test were all set at 5% level of significance and the acceptance proportion of each candidate distributions for both the trimmed and untrimmed samples are computed.

In model selection, the candidate distribution is often arbitrarily chosen from a set of commonly used probability distribution functions (PDFs). Therefore, the evaluation of a model selection method should consider not only the accuracy of identifying the correct distribution when the sample is drawn from the true parent distribution, but also its ability to reject incorrect hypotheses when the assumed PDF differs from the actual one. An ideal model selection method should maintain high accuracy when the data are generated from the assumed PDF while also minimizing the rate of false acceptance when the data originate from other distributions. To address this, a formula is proposed to describe the power of a model selection method for a specific PDF:

$$Power(MSH_(b_{i,j})) = PMSH_(b_{i,s=h}) \times [\sum_(s \neq h) (1 - PMSH_(b_{i,j}))] / 4 \tag{6}$$

Considering the powers for different probability density functions, the comprehensive power for each model selection method can be described as follows:

$$Power(M) = [\sum_s Power(MS)] / 4. \tag{7}$$

The terms in Equations (6) and (7) are as defined in Equation (5).

RESULTS AND DISCUSSION

Figure 1 to Figure 12 gives the acceptance proportion of the four model selection methods for samples generated from the Lognormal, Gumbel, and P3 parent distributions for trimmed and untrimmed samples for varying sample and subsample size.

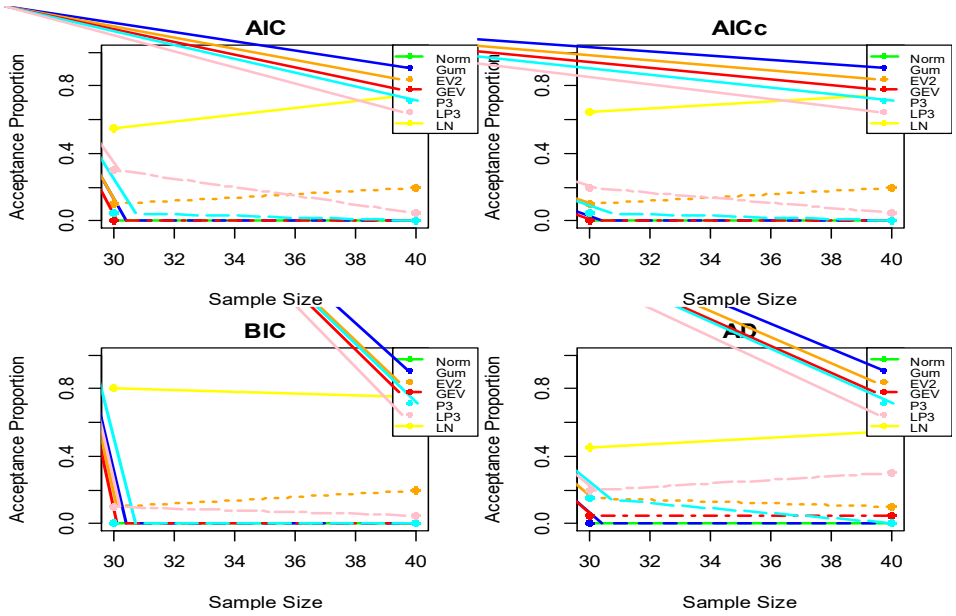

Figure 1: The acceptance proportions of the model selection methods for a subsample size of 20, based on untrimmed samples generated from the lognormal parent distribution.

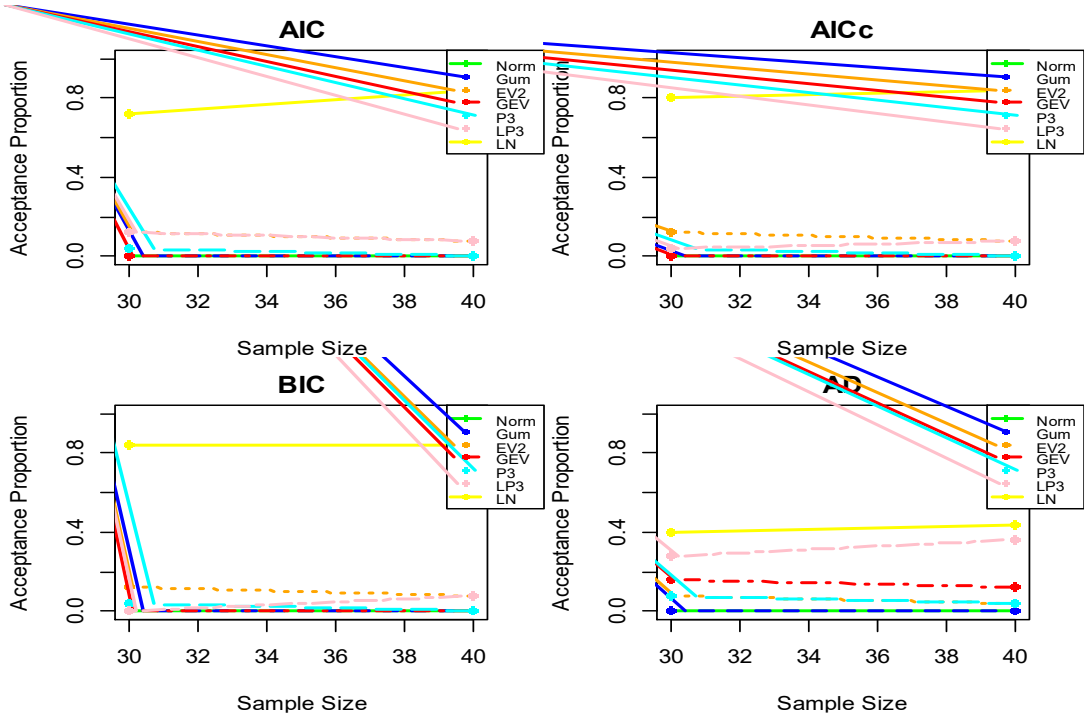

Figure 2: The acceptance proportions of the model selection methods for a subsample size of 25, based on untrimmed samples generated from the lognormal parent distribution.

Figures 1 and 2 displays the acceptance proportions of the model selection methods for the candidate distributions when the true parent distribution is Lognormal, based on untrimmed samples with varying sample and subsample sizes. The model selection methods correctly identify the lognormal distribution as the underlying distribution in approximately 0.40 to 0.84 of times. Among the methods, the Bayesian Information Criterion (BIC) yields the best performance in recognizing the true distribution, followed by AICc, AIC, and the Anderson-Darling (AD) test. Although AIC and AICc exhibit similar selection tendencies, AICc generally shows superior performance relative to AIC. The AD test successfully chooses the lognormal distribution in approximately 0.40 to 0.55 of the cases.

An analysis of Figures 3 and 4 reveals that, for untrimmed samples, the model selection methods sometimes struggle to differentiate the Gumbel distribution from the lognormal distribution. As shown in Figure 3, only the BIC consistently selects the underlying distribution, with an acceptance proportion of approximately 0.40. However, as the sample size increases, the Gumbel distribution is identified as the parent distribution by each model selection methods in approximately 0.33 to 0.50 of times across the entire set of subsample sizes. Among the methods, the BIC demonstrate the highest effectiveness at recognizing the underlying distribution, followed by the AICc, AIC, and Anderson-Darling (AD) test, in that order.

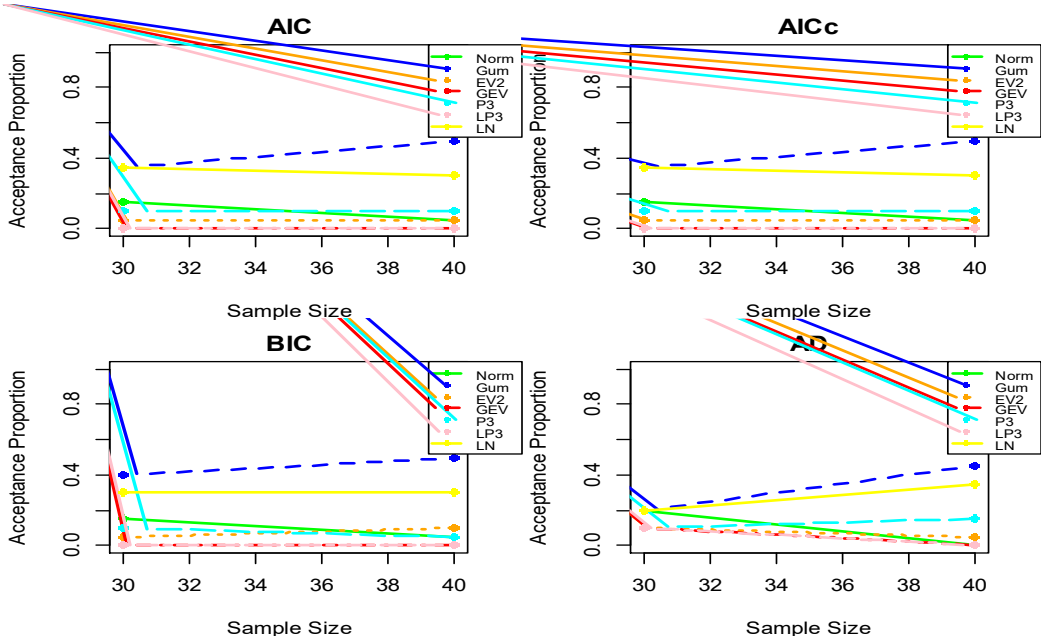

Figure 3: The acceptance proportions of the model selection methods for a subsample size of 20, based on untrimmed samples generated from the Gumbel parent distribution.

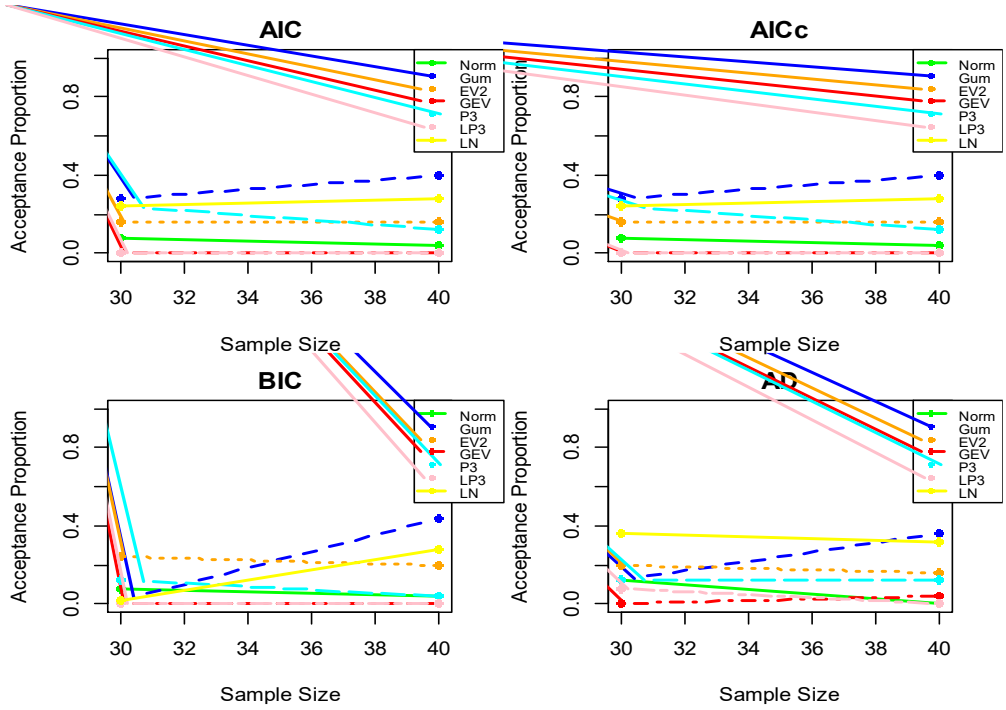

Figure 4: The acceptance proportions of the model selection methods for a subsample size of 25, based on untrimmed samples generated from the Gumbel parent distribution.

When the Pearson Type III (P3) distribution is used as the parent distribution, the model selection methods tend to favor selecting a two-parameter distribution over the true three-parameter parent distribution across all sample and subsample sizes for the untrimmed data (see Figures 5 and 6). This tendency may reflect an implicit preference for model parsimony which favours the simplest model that sufficiently models the inherent data generation.

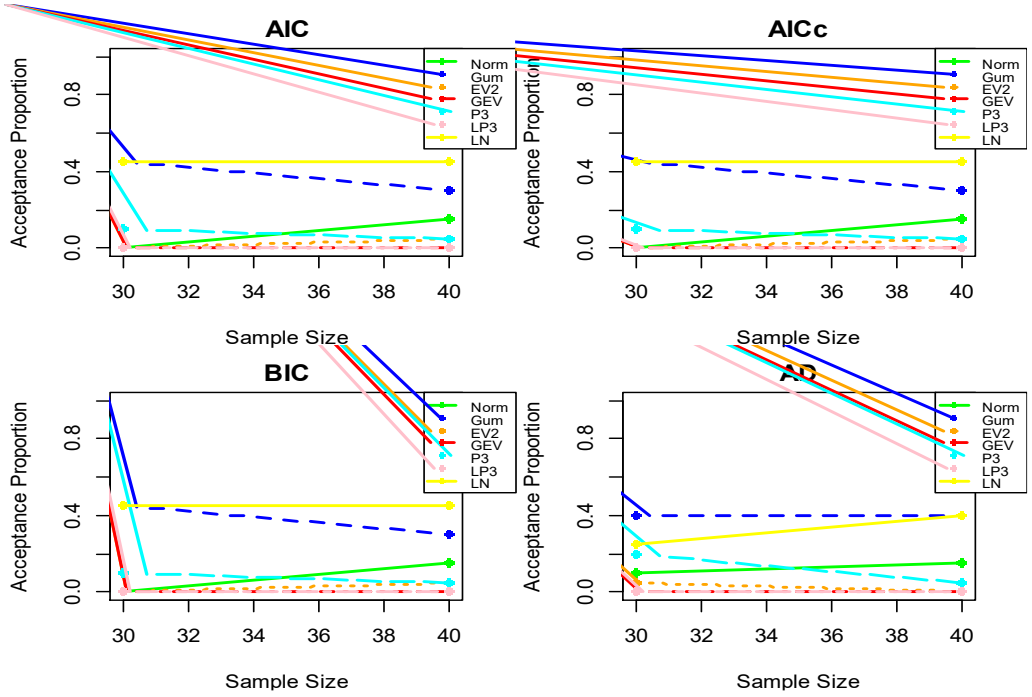

Figure 5: The acceptance proportions of the model selection methods for a subsample size of 20, based on untrimmed samples generated from the P3 parent distribution.

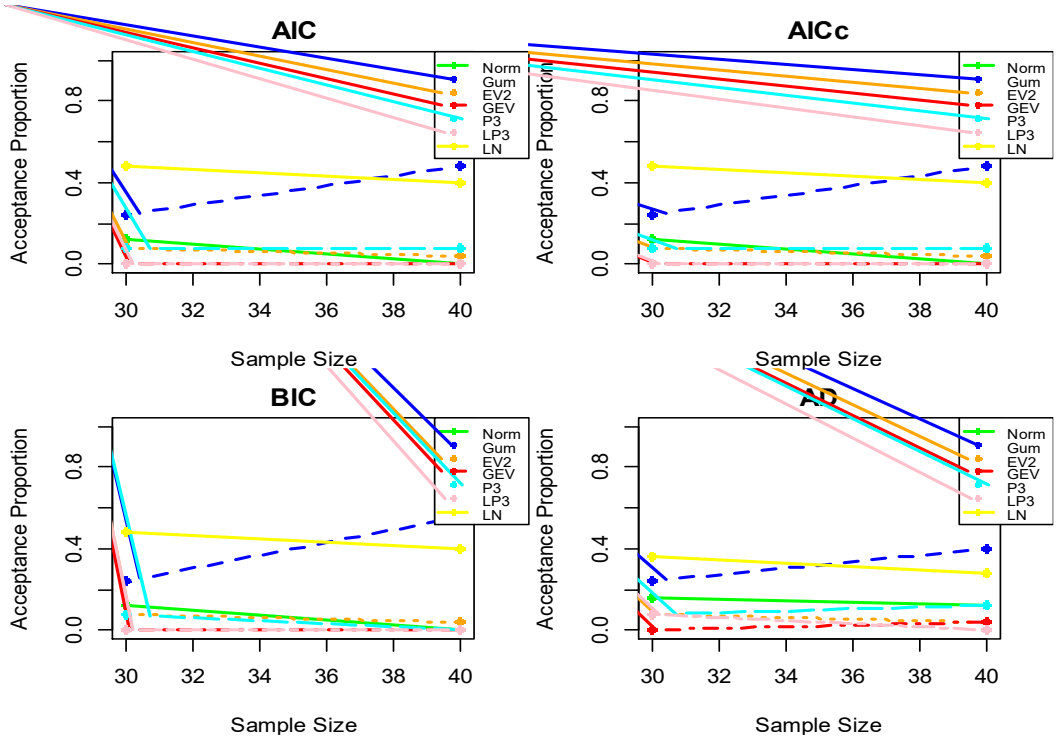

Figure 6: The acceptance proportion of model selection methods for a subsample size of 25, based on untrimmed samples generated from the P3 parent distribution.

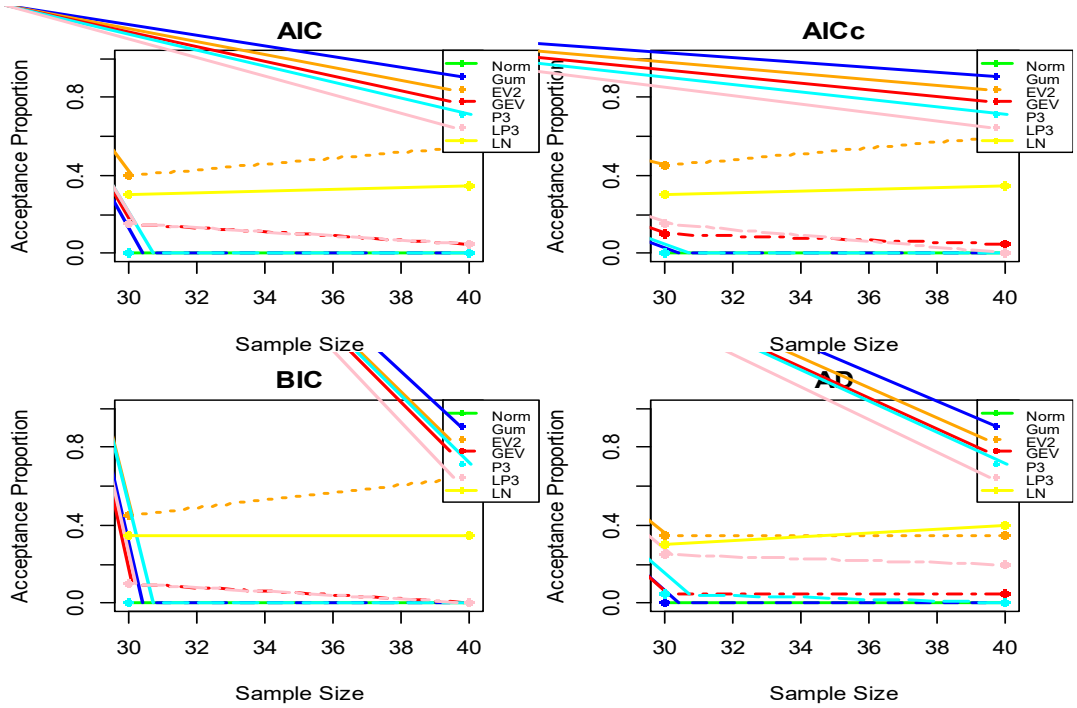

Figure 7: The acceptance proportions of the model selection methods for a sub- sample size of 20 with a 10% trimming proportion, based on samples generated from the Lognormal parent distribution.

As shown in Figure 7, when the lognormal distribution is the parent distribution for trimmed samples, the findings reveal that either the lognormal or the EV2 distribution is selected as the underlying distribution at a 10% trimming level. The lognormal distribution is preferred in approximately 0.30-0.48 of cases, while the EV2 distribution is chosen in about 0.28-0.45 of cases. However, increasing the trimming level to 15% for a sample size of 40, all model selection methods identify the EV2 distribution as the underlying distribution in approximately 0.57-0.67 of cases (see Figure 8).

The tendency of the four model selection criteria to identify the EV2, which is a three-parameter distribution as the parent distribution can be explained to be the influence of lower observed values favoring the selection of three-parameter models.

When the Gumbel distribution is used as the parent distribution and trimming is introduced, all model selection criteria consistently identify the P3 distribution as the best-fitting model (see Figure 9). This suggests a chance for the selection techniques to favor a three-parameter distribution over the actual parent distribution having two-parameter. However, as one of the objectives of extreme value analysis is the accurate estimation of quantiles, this outcome may not invariably pose a limitation. In certain situations, a model that fits the data well might prove not be optimal for quantile estimation. This shifts the focus from identifying the true parent distribution to selecting the most operationally effective model for quantile estimation.

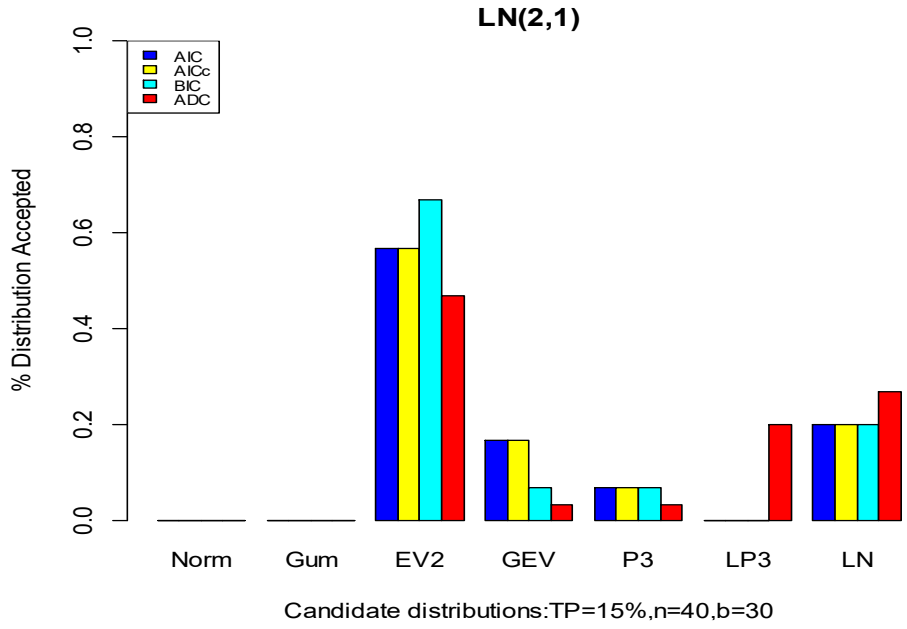

Figure 8: The acceptance proportions of the model selection methods for $n = 40$, $b = 30$, and a 15% TP, based on samples generated from the LN parent distribution.

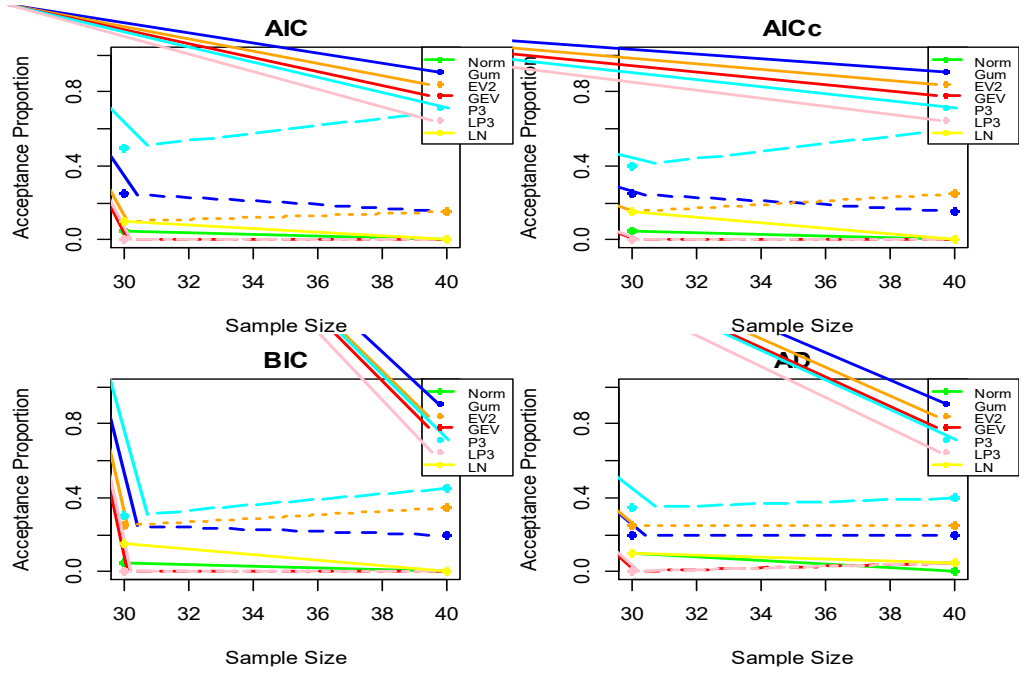

Figure 9: The acceptance proportions of the model selection methods for a sub- sample size of 20 with a 10% trimming proportion, based on samples generated from the Gumbel parent distribution.

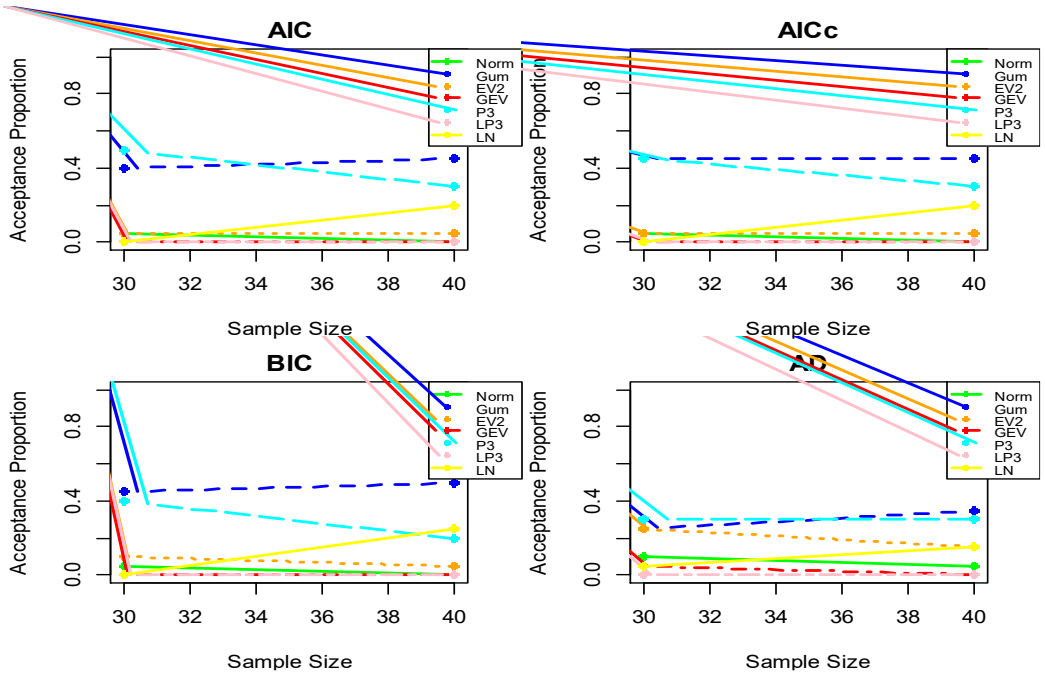

Figure 10: The acceptance proportions of the model selection methods for a sub- sample size of 20 with a 10% trimming proportion, based on samples generated from the P3 parent distribution.

By analyzing Figures 10, 11, and 12, it can be observed that when trimming is introduced and the Pearson Type III distribution serves as the parent distribution, each model selection methods accurately identify it across all sample and subsample sizes. Among the methods, the AIC consistently performs best, regardless of variations in sample size, subsample size, or trimming proportion.

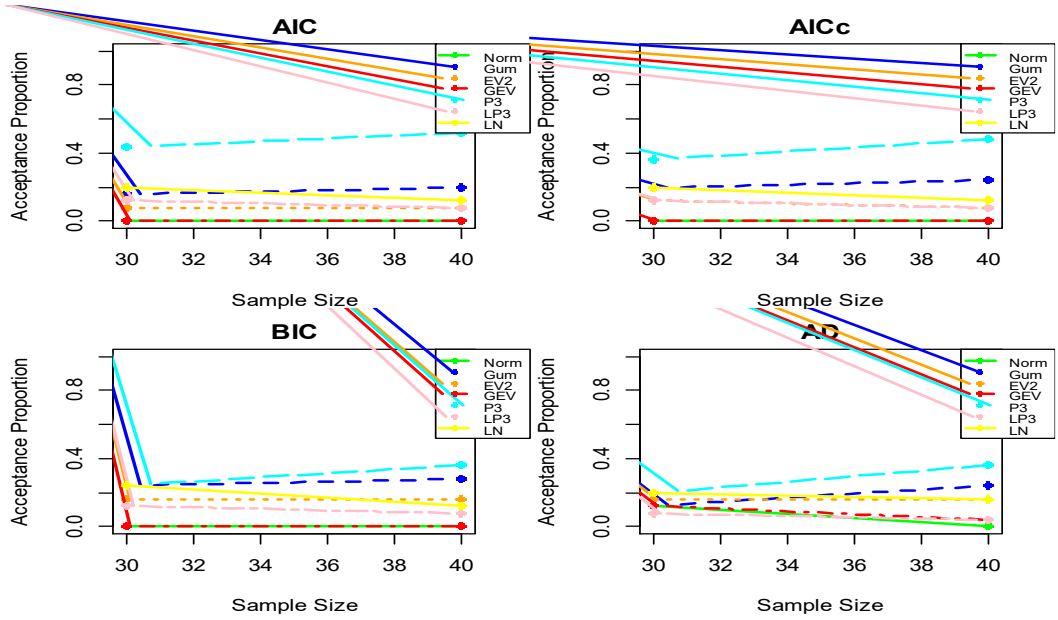

Figure 11: The acceptance proportions of the model selection methods for a sub- sample size of 25 with a 15% trimming proportion, based on samples generated from the P3 parent distribution.

The ability of the model selection methods to correctly identify the true underlying distribution following trimming can be attributed to the removal of the undesirable influence of lower observations on the upper tail of the distribution, particularly in the case of three-parameter distributions.

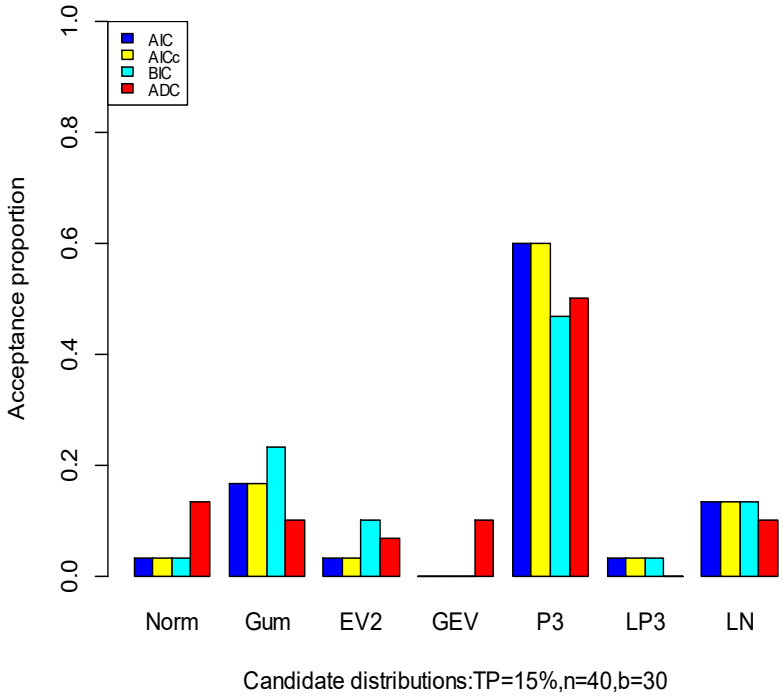

Figure 12: The acceptance proportions of the model selection methods for $n = 40$, $b = 30$, and a 15% TP, based on samples generated from the P3 parent distribution.

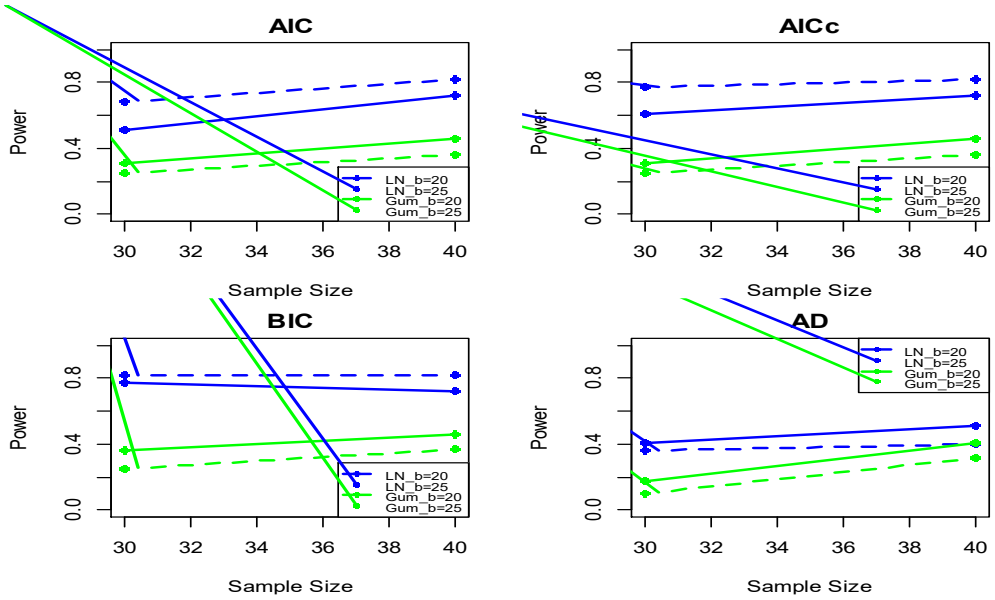

Figure 13: The power of the model selection criteria for the two-parameter Lognormal and Gumbel parent distributions.

Figure 13 presents the statistical power of the AIC, AICc, BIC, and Anderson- Darling (AD) test when the parent distributions are the two-parameter Lognormal and Gumbel distributions, respectively. The power generally increases as both the sample size and subsample size increase, for all model selection methods, except for BIC, which exhibits a somewhat different behavior when the parent distribution is Lognormal. When the parent distribution is Gumbel, higher power is observed at a subsample size of 20. In contrast, for the lognormal distribution, higher power is observed at a subsample size of 25, with the exception of the AD test, which achieves greater power at a subsample size of 20. As the sample size increases, the differences in power across subsample sizes become less pronounced for all model selection methods.

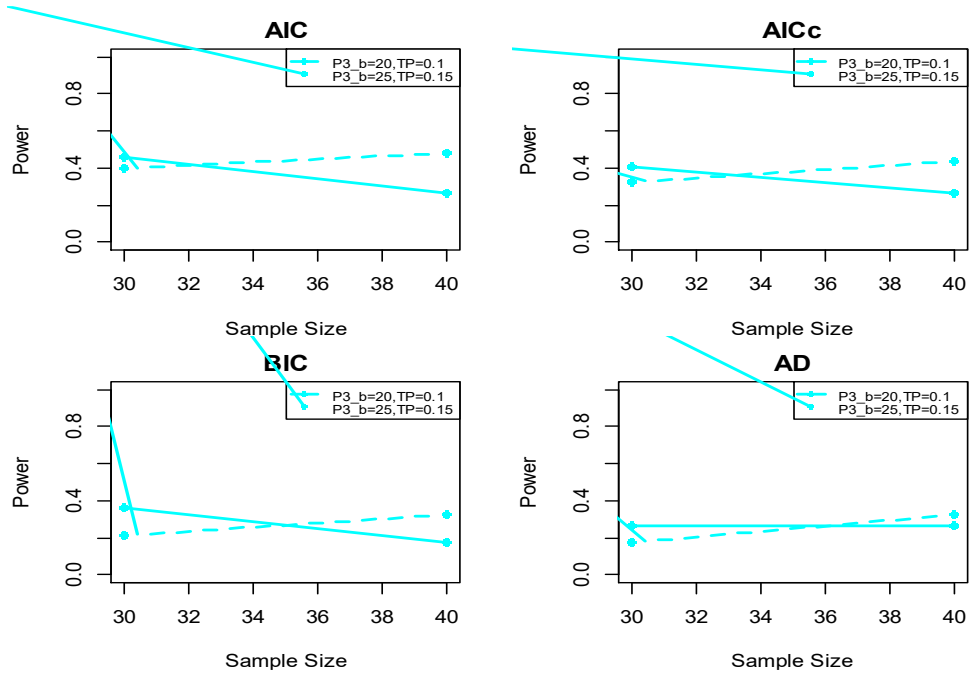

Figure 14: The power of the model selection criteria for the three-parameter Pearson Type III (P3) parent distribution.

Figure 14 displays the statistical power of the AIC, AICc, BIC, and Anderson- Darling (AD) test when the parent distribution is the Pearson Type III (P3) distribution. The power increases with both the length of the series and the subsample size for each test method, with a trimming proportion of 15% resulting in higher power. When the subsample size increases to 30, the AIC and AICc methods exhibit identical power, followed by the AD test and then the BIC (see Figure 15).

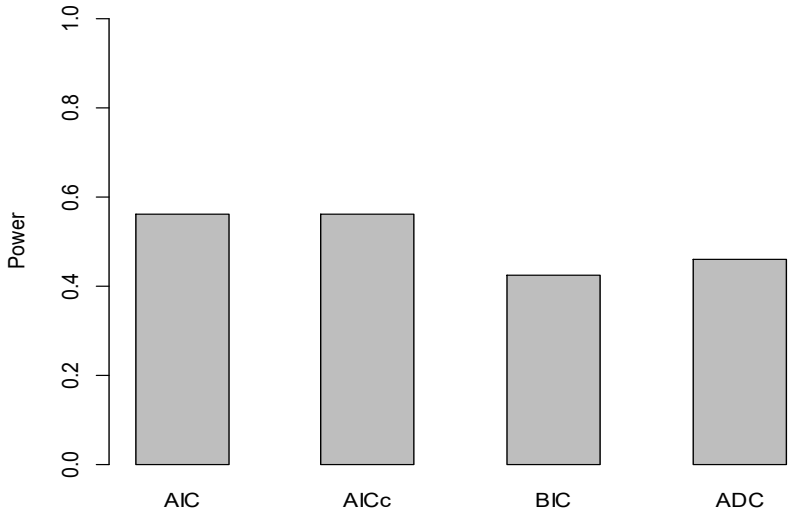

Figure 15: The power of model selection criteria for $n = 40$, $b = 30$, and a 15% TP, based on samples generated from the P3 parent distribution.

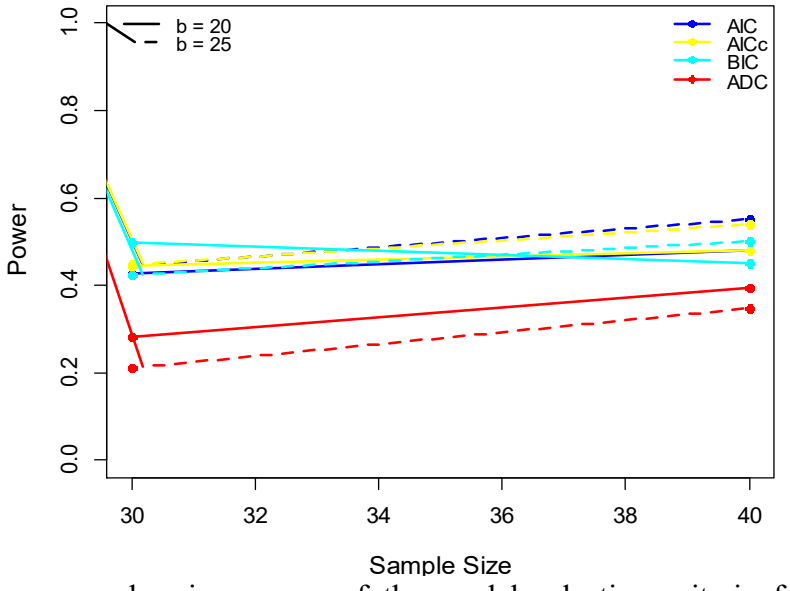

Figure 16: The comprehensive power of the model selection criteria for the Lognormal, Gumbel, and P3 parent distributions

Figure 16 shows that the overall power of all model selection methods in- creases with the length of the series, except for the BIC. Among the methods, AIC demonstrates the highest overall power, followed by AICc, BIC, and the Anderson-Darling (AD) test. The differences in power among AIC, AICc, and BIC are relatively small.

CONCLUSION

An intensive numerical analysis was conducted to evaluate the performance of four model selection criteria: the Akaike Information Criterion (AIC), the corrected Akaike Information Criterion (AICc), the Bayesian Information Criterion (BIC), and the Anderson-Darling (AD) test, in the presence of subsampling and trimming. The parent distributions considered were the Lognormal, Gumbel, and Pearson Type III (P3) distributions. Based on the comparison of the model selection methods, several conclusions can be drawn: No single model selection method consistently outperforms the others across all cases. The model selection methods supported by subsampling successfully identify the true underlying distribution for untrimmed samples when the distribution is two-parameter. However, this success is not as evident when the parent distribution is three-parameter. When trimming is applied, all model selection methods successfully identify the true parent distribution even for three-parameter distributions. Overall, the combination of trimming and subsampling, together with model selection method, produces favourable results for extreme value analysis. The most powerful model selection methods are BIC for the Lognormal and Gumbel distribution, and AIC for the P3 distribution. Regarding comprehensive power, the AIC exhibits the best overall performance, followed by AICc, BIC, and the Anderson-Darling test. This study highlights the viability of trimming and subsampling in strengthening the power of model selection, which plays a central role in extreme value modelling especially when sample sizes are limited and distributions are asymmetric.

ACKNOWLEDGEMENT

We extend our gratitude to Universiti Putra Malaysia for their financial backing of this research through the PUTRA GRANT GP/2023/9753100.

REFERENCES

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike* (pp. 199–213). New York, NY: Springer New York.

Beirlant, J., & Bladt, M. (2025). Tail classification using non-linear regression on model plots. *Extremes*, 1–25.

Bhattarai, K. P. (2004). Partial L-moments for the analysis of censored flood samples/Utilisation des L-moments partiels pour l'analyse d'échantillons tronqués de crues. *Hydrological Sciences Journal*, **49(5)**.

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, **7(6)**: 679–692.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33(2)**: 261–304.

Calenda, G., Mancini, C. P., & Volpi, E. (2009). Selection of the probabilistic model of extreme floods: The case of the River Tiber in Rome. *Journal of Hydrology*, **371(1–4)**: 1–11.

Chan, S., Chu, J., Zhang, Y., & Nadarajah, S. (2022). An extreme value analysis of the tail relationships between returns and volumes for high frequency crypto currencies. *Research in International Business and Finance*, **59**: 101541.

Chernick, M. R. (2011). *Bootstrap methods: A guide for practitioners and researchers*. John Wiley & Sons.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208, p. 208). London: Springer.

Cunnane, C. (1989). Statistical distribution for flood frequency analysis. *WMO Operational Hydrology*, Report No. 33, WMO-No. 718, Geneva, Switzerland.

Das, S. (2017). An assessment of using subsampling method in selection of a flood frequency distribution. *Stochastic Environmental Research and Risk Assessment*, **31**: 2033–2045.

Di Baldassarre, G., Laio, F., & Montanari, A. (2009). Design flood estimation using model selection criteria. *Physics and Chemistry of the Earth, Parts ABC*, **34(10–12)**: 606–611.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution* (pp. 569–593). New York, NY: Springer New York.

Flammini, A., Dari, J., Corradini, C., Saltalippi, C., & Morbidelli, R. (2022). Areal reduction factor estimate for extreme rainfall events. In *Rainfall* (pp. 285–306). Elsevier.

Gado, T. A. (2016). An at-site flood estimation method in the context of nonstationarity I. A simulation study. *Journal of Hydrology*, **535**: 710–721.

Haddad, K., & Rahman, A. (2011). Selection of the best fit flood frequency distribution and parameter estimation procedure: a case study for Tasmania in Australia. *Stochastic Environmental Research and Risk Assessment*, **25**: 415–428.

Kobierska, F., Engeland, K., & Thorarinsdottir, T. (2018). Evaluation of design flood estimates: a case study for Norway. *Hydrology Research*, **49(2)**: 450–465.

Kousar, S., Khan, A. R., Ul Hassan, M., Noreen, Z., & Bhatti, S. H. (2020). Some best-fit probability distributions for at-site flood frequency analysis of the Ume River. *Journal of Flood Risk Management*, **13(3)**: e12640.

Laio, F., Di Baldassarre, G., & Montanari, A. (2009). Model selection techniques for the frequency analysis of hydrological extremes. *Water Resources Research*, **45(7)**.

Linhart, H., & Zucchini, W. (1986). *Model selection*. John Wiley & Sons.

Pan, X., & Rahman, A. (2022). Comparison of annual maximum and peaks-over-threshold methods with automated threshold selection in flood frequency analysis: a case study for Australia. *Natural Hazards*, 1–26.

Politis, D. N., Romano, J. P., & Wolf, M. (1999). Subsampling in the IID Case (pp. 39–64). Springer New York.

Reghenzani, F., Massari, G., Santinelli, L., & Fornaciari, W. (2019). Statistical power estimation dataset for external validation GoF tests on EVT distribution. *Data in Brief*, **25**: 104071.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Wang, Q. J. (1996). Using partial probability weighted moments to fit the extreme value distributions to censored samples. *Water Resources Research*, **32(6)**: 1767–1771.

Zeng, X., Wang, D., & Wu, J. (2015). Evaluating the three methods of goodness of fit test for frequency analysis. *Journal of Risk Analysis and Crisis Response*, **5(3)**: 178–187.